# 2025 Connecticut Sports Analytics Data Challenge

## Swinging for the Win: Unraveling the Impact of Swing Decisions on Run Expectancy & Win Probability

Team DBD

Brett Cerenzio, Danielle Napierski, and Dan Griffiths

Syracuse University

## Abstract

For some time, there has been a stark disagreement among fans about what the ideal swing type is for batters. Some fans argue that batters are too selfish in today's game, citing the low batting averages, rising strikeout rates, and the increasing amount of "home run" swings players have become more accustomed to taking. Others argue that low batting averages and rising strikeout rates have to do more with the advancement of pitching and that hitting for power, rather than contact, typically yields better results in the long run. There are also those who argue that batters should change their swings in different situations. For example, they can swing for the fences when the bases are empty, but should look to shorten their swing and emphasize contact when there are runners in scoring position.

For this project, we wanted to evaluate if different swing types are needed to maximize run production or if there's one swing type that provides a perfect balance between power and contact. We utilized k-means clustering to classify seven different swing types and then created two RBI matrices. One evaluates the expected number of RBIs on a given swing based on swing type, base situation, and number of outs, while the other finds the probability of at least 1 RBI scoring on a given swing based on swing type, base situation, and number of outs. We also created an XGBoost model that predicted the change in win probability based on swing type, base situation, number of outs, count, score, inning, and batting team score differential.

## Introduction

The initiation of swing metrics to Statcast has opened the door to many new possibilities that can change how we view hitting. Just this year, Statcast introduced two new metrics that utilize swing metrics: Squared Up% and Blast%. **Squared Up%** measures the quality of contact made during a swing by comparing the ball's actual exit velocity to the maximum possible exit velocity for that swing. A swing is considered "squared up" if the exit velocity exceeds 80% of the maximum possible exit velocity, indicating solid and efficient contact. **Blast%** identifies swings that achieve elite-level power and quality of contact by combining bat speed and the percent squared up, where percent squared up equals exit velocity divided by max possible exit velocity. If the average of (percent squared up)*100 and bat speed is greater than or equal to 82, then the result is a blast.

We wanted to look into how different types of swings can impact run scoring. There's a long-standing argument among many in the baseball industry about when it's appropriate to shorten one's swing. On one hand, a shorter swing typically results in a higher contact rate, but it also severely limits the amount of bat speed a hitter can accumulate, which limits the amount of damage a batter can do on a single swing. This can be seen when a batter shortens his swing with

2 strikes to decrease his likelihood of striking out. In this paper, we are going to classify each swing type and identify which swings are best suited for different situations.

## Data

Using pitch-by-pitch data from the 2024 regular season provided by Statcast, we refined the variables for our analysis to focus on key metrics. These include bat speed and swing length—two newly introduced metrics—along with pitch location, batted ball type (ground ball, line drive, or fly ball), pitch result (e.g., swinging strike, hit into play), wOBA (weighted on-base average), batting team score differential (difference between the batting and fielding teams' scores), and delta win expectancy. Additionally, contextual factors such as pitch count, number of outs, and inning were included to analyze various scenarios where choking up might be advantageous.

We also gathered play-by-play data from MLB Stats API using the baseballr package in RStudio. This dataset provided additional details, such as the number of RBIs (runs batted in) recorded during each plate appearance, which we used to calculate RBIs for individual swings. Another key feature of this data was the summarized base situations—categorized as empty, runners on base, runners in scoring position (RISP), or bases loaded—offering valuable context for our preliminary analysis.

We combined the two datasets by creating a shared game_pitch_id variable, which represents the cumulative number of pitches thrown at a given point in the game. For the Statcast data, this was achieved by first ordering by at-bat number (where a value of 1 represents a batter's first appearance in the game) and then by pitch number, which tracks the number of pitches for each plate appearance. For the MLB Stats API data, we utilized its about.atBatIndex variable, which corresponds to the at-bat number in Statcast, and the start time of each pitch to order the data. Finally, we merged the two datasets using game_pk (a unique identifier for each game) and the newly created game_pitch_id.

Furthermore, using the new bat speed and swing length variables, we were able to calculate two additional metrics using physics, along with some help from Driveline Baseball:

**Rotational Acceleration:** how quickly a batter accelerates his bat into the swing plane.

$$Rotational\ Acceleration = ((Bat\ Speed * 1.46667)^2/(2 * Swing\ Length))/32.174$$

*Where 1.46667 is the conversion of MPH to ft/s and 32.174 is the conversion from* $ft/s^2$ *to g-force, which is the industry standard measurement for rotational acceleration.*

**Time to Contact:** the duration between the beginning of the downswing and the moment of impact with the ball.

$$Time\ to\ Contact\ = (2 * Swing\ Length) / (Bat\ Speed\ * \ 1.46667)$$

*Where 1.46667 is the conversion of MPH to ft/s.*

To wrap up our data manipulation, we created a few binary variables. One was for if contact was made or if it was a whiff. Another was if it was a **competitive swing**, which Statcast defines as a swing where the bat speed is either above the batter's 10th percentile bat speed or exceeds 60 MPH with a launch speed greater than 90 MPH, indicating a meaningful and impactful swing. Lastly, we calculated Squared Up% and Blast% to evaluate the quality of contact of each swing.

## Methodology

To get a better understanding of the data before jumping to larger analysis, we first wanted to see how different bat speeds and swing lengths impacted various metrics and how players change their swings depending on the situation.

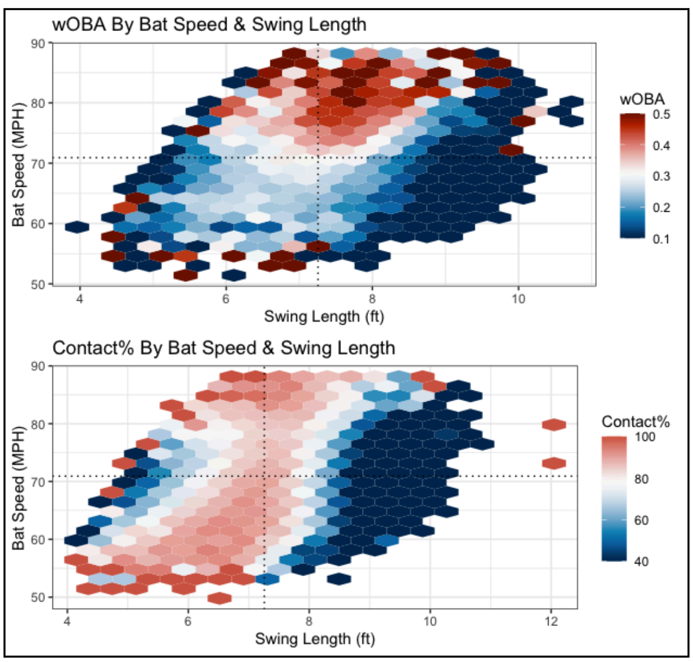



Figure 1: wOBA and Contact% based on bat speed and swing length. Faster and longer swings tend to result in an increase in wOBA but at the expense of contact rate.

In Figure 1, we noticed that wOBA tends to increase with higher bat speeds and longer swing lengths. However, this also led to lower contact rates, which can be detrimental with runners on base.

Also, batters tend to adjust their swings strategically based on game situations, such as base runners, outs, and pitch counts, to maximize their team's scoring opportunities. For example, if there is a runner on third and less than two outs, we know that hitting towards the right side of the field would be optimal because this could score the runner on third. Batters adjust their swing so that it is more likely for their teammate to score. Their swing can also change based on the number of outs. Back to that example before, imagine there were now two outs. Even if the batter adjusts his swing and hits a hard ground ball on the edge of the grass, there is still a good chance of getting the out at first base. Another big reason batters adjust their swings is when they have to protect the plate with two strikes. This is when batters often choke up on the bat, which will affect bat speed and swing length. The real question is, should a batter make these changes to his swing?

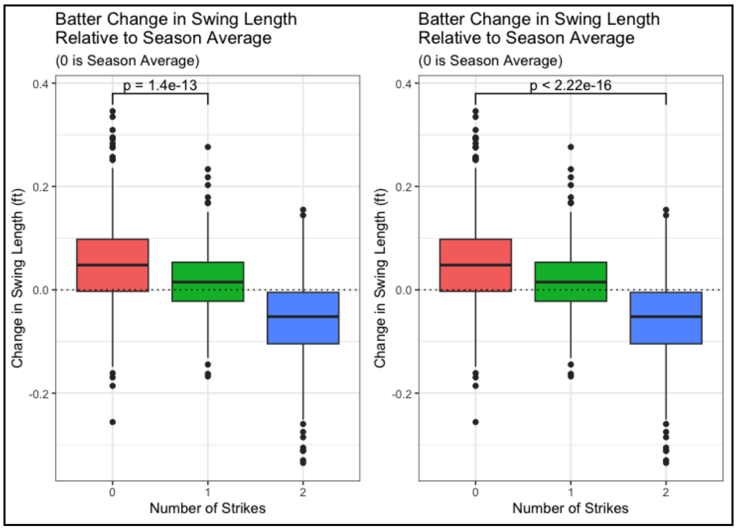



Figure 2: Plot showing change in swing length based on the number of strikes in the count (Where the batter season average swing length = 0). The plot on the left shows there's statistical significance in the difference in swing length in 0 and 1 strike counts and the plot on the right shows there's statistical significance between swing length in 0 and 2 strike counts. This shows that batters do intentionally decrease their swing length as strikes increase for the sake of contact.

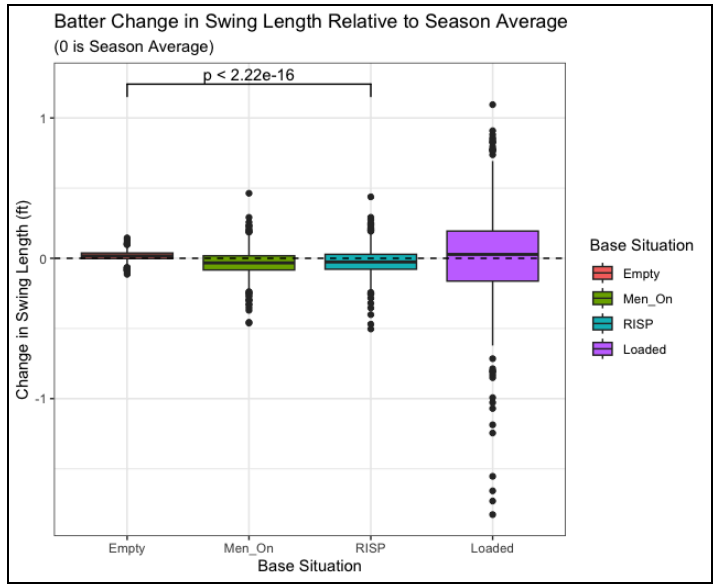



Figure 3: Boxplot comparing individual batter swing lengths in different base situations (where the batter's season average swing length = 0). Here, we can see that batters decrease their swing length with men on and with RISP to increase contact%.

To start, we applied k-means clustering to classify different swing types based on bat speed, swing length, rotational acceleration, and time to contact in our filtered data frame containing competitive swings only. As mentioned earlier, competitive swings are swings in the top 90% of any batter's swings as well as any 60+ MPH swings that result in an exit velocity of 90+ MPH.

We used the elbow method to find the optimal number of clusters, which ended up being seven different swing types. Every single pitch was assigned a cluster, but to look at which cluster each batter belonged to, we addressed which swing type they used the most. Note that some players did have a tie between the number of appearances in clusters, but they didn't face a lot of pitches throughout the season. In addition, note that Table 1 below represents the average of every swing in the cluster, not the season averages of players in the cluster.

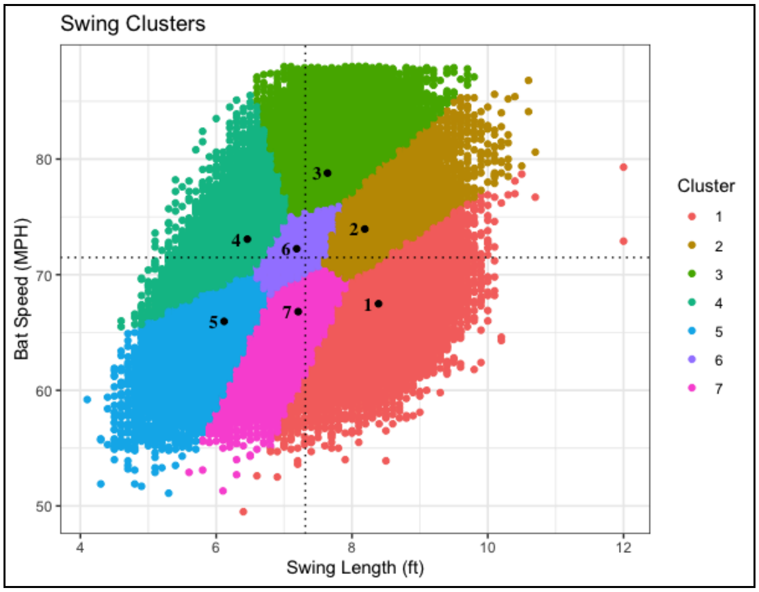



Figure 4: Plot showing the different swing clusters created by k-means clustering.

| Cluster | # of Players | Bat Speed | Swing Length | Rot. Accel. | Time to Contact | Squared Up Swing% | Blast Swing% | Barrel Swing% | wOBA | Whiff% |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 18 | 67.5 | 8.4 | 18.2 | 0.170 | 10.06% | 0.94% | 0.35% | 0.148 | 48.43% |
| 2 | 100 | 73.9 | 8.2 | 22.4 | 0.151 | 16.89% | 7.09% | 2.88% | 0.301 | 32.53% |
| 3 | 66 | 78.8 | 7.6 | 27.2 | 0.132 | 27.64% | 18.98% | 8.15% | 0.443 | 15.80% |
| 4 | 71 | 73.1 | 6.5 | 27.7 | 0.121 | 22.86% | 9.36% | 2.80% | 0.301 | 22.30% |
| 5 | 92 | 66.0 | 6.1 | 23.8 | 0.126 | 22.10% | 2.13% | 0.53% | 0.250 | 16.56% |
| 6 | 237 | 72.2 | 7.2 | 24.3 | 0.136 | 31.82% | 12.16% | 4.27% | 0.350 | 12.45% |
| 7 | 92 | 66.8 | 7.2 | 20.7 | 0.147 | 28.84% | 3.94% | 1.22% | 0.260 | 10.59% |

Table 1: Summary statistics for each individual cluster.

**Cluster 1:** Swings that are both slow and long. This isn't many players' primary swing, and can infer that this swing occurs most often when they're fooled on a pitch.

**Cluster 2**: Swings that are fast and long. These players swing for the fences, often at the expense of contact.

**Cluster 3:** These swings achieve high rotational acceleration by combining fast swing speeds with league-average swing length.

**Cluster 4:** These swings are similar to Cluster 3, however, they combine league-average swing speeds with shorter swing lengths.

**Cluster 5:** Swings that are slow and short. These value being quick to the ball much more than maximizing exit velocity.

**Cluster 6:** Swings that are both average speed and length. These can be considered "balanced swings."

**Cluster 7:** Swings that are slow by average length. These swings look to maximize contact by allowing for adjustability throughout the swing.

**Notable Batters**

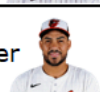
| | | | |
|---|---|---|---|
| Cluster 1 | Jose Altuve  | Isaac Paredes  | Jeimer Candelario  |
| Cluster 2 | Aaron Judge  | Teoscar Hernández  | Willy Adames  |
| Cluster 3 | Giancarlo Stanton  | Shohei Ohtani  | Vladimir Guerrero Jr.  |
| Cluster 4 | Bryce Harper  | Bobby Witt Jr.  | Corey Seager  |
| Cluster 5 | Luis Arráez  | Justin Turner  | Steven Kwan  |
| Cluster 6 | Anthony Santander  | Jackson Chourio  | José Ramírez  |
| Cluster 7 | Santiago Espinal  | Nicky Lopez  | Whit Merrifield  |

Table 2: Displaying notable batters who appeared a lot in these clusters. Note that these batters also had swings in other clusters, but are listed in the cluster they had the highest percentage of swings in. For example, Aaron Judge had 44 swings in Cluster 1, 478 swings in Cluster 2, 410 in Cluster 3, 40 in Cluster 4, 1 in Cluster 5, 146 in Cluster 6, and 1 in Cluster 7. Since most of his swings occurred in Cluster 2, we classified him as Cluster 2.

Using these clusters, we then created an RBI expectancy matrix based on the base situation and number of outs for each cluster. We also created a matrix that found the probability of at least one RBI scoring in a given base situation and out scenario for each cluster. Both the expectancies and probabilities are on a per-swing basis.

We then ran an extreme gradient boosting (XGBoost) model to predict the change in win probability based on various factors, including swing type, base situation, outs, strikes, balls, inning, and the score difference between the batting and fielding teams (e.g., a score difference of -1 if the batting team is losing by 1, or a difference of 2 if the batting team is leading by 2). This model aims to help us determine the most effective strategies in late, close-game situations.

## Results

Following this methodology, we were able to assess the impact of each swing type on run production and game outcomes. The seven swing types revealed unique characteristics and situational strengths, with Cluster 3 and Cluster 6 emerging as the most impactful swing types in determining changes in win probability (ΔWP).

**Cluster 1**: With a wOBA of .148 and a Whiff% of 48.43%, this cluster proved to be the least effective in terms of run production. It performed significantly worse than other clusters in all situations.

**Cluster 2**: This cluster recorded a Blast Swing% of 7.09% and a Barrel Swing% of 2.88%. However, its Whiff% (32.53%) limited its effectiveness in scenarios requiring consistent contact, making it more suitable for early-inning opportunities when the goal is to establish a lead.

**Cluster 3:** This cluster stood out as the most influential swing type, combining high rotational acceleration (27.2 g) with a balanced mix of fast swing speed (78.8 mph) and league-average swing length (7.6 ft). It achieved a wOBA of .443, the highest among all clusters, and excelled in producing high-quality hits, as evidenced by its Blast Swing% (19%) and Barrel Swing% (8.2%). Cluster 3 also recorded the lowest ground ball percentage (35.5%), making it highly effective in generating extra-base hits and sustaining rallies. This cluster was particularly impactful in late-game scenarios, maximizing ΔWP during the seventh inning or later with a score differential of one run or less.

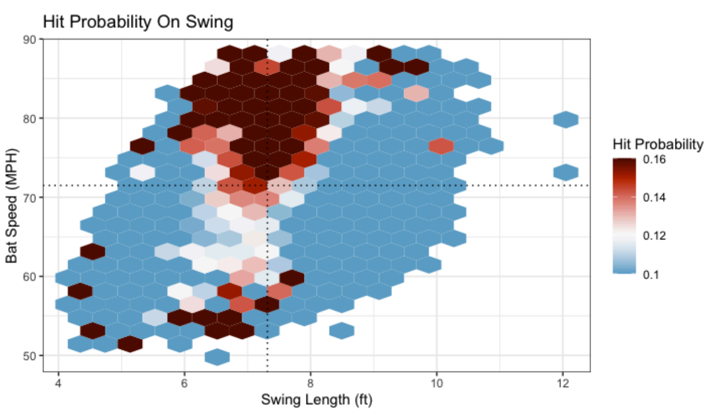



Figure 5: Plot showing the probability of a swing resulting in a hit based on bat speed in swing length. Notice how the swings that give the batter the best chance of recording a hit on the swing highly correlate with swings in Cluster 3.

There were 57 players that had Cluster 3 as their second-highest appearance, and a few players missed that cut by a relatively low number of swings, including Bryce Harper (Cluster 4), Aaron Judge (Cluster 2), Anthony Santander (Cluster 6), Bobby Witt Jr. (Cluster 4), Riley Greene (Cluster 4), and Jackson Chourio (Cluster 6). In Figure 6, we can see that there wasn't a large difference in the number swings between Cluster 3 and their main cluster.

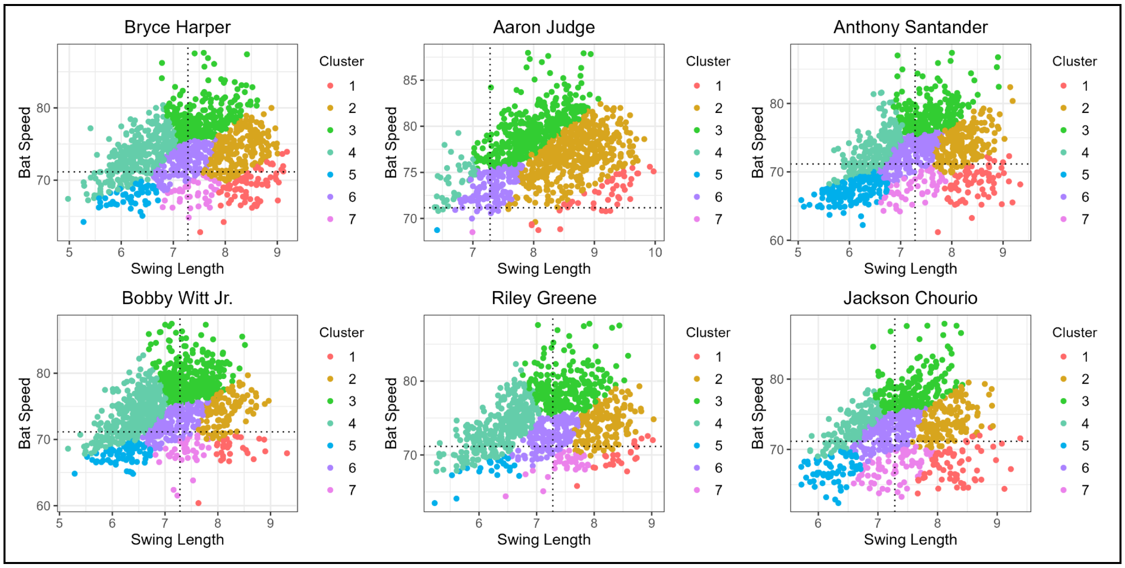



Figure 6: Each plot illustrates every competitive swing of the six batters that fell short of making it to Cluster 3. Lime green represents Cluster 3, gold represents Cluster 2, aquamarine green represents Cluster 4, and purple represents Cluster 6.

| Batter | Cluster | Bat Speed | Swing Length | Rot. Accel. | Time to Contact | BA | OBP | OPS | RBI | HR | wOBA | Whiff% |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bryce Harper | 4 | 74.1 | 7.3 | 25.5 | 0.134 | 0.285 | 0.373 | 0.898 | 87 | 30 | 0.397 | 24.60% |
| Aaron Judge | 2 | 77.2 | 8.2 | 24.5 | 0.145 | 0.322 | 0.458 | 1.159 | 144 | 58 | 0.483 | 28.93% |
| Anthony Santander | 6 | 73.1 | 7.2 | 24.9 | 0.135 | 0.235 | 0.308 | 0.814 | 102 | 44 | 0.344 | 17.83% |
| Bobby Witt Jr. | 4 | 74.7 | 7.1 | 26.5 | 0.130 | 0.332 | 0.389 | 0.977 | 109 | 32 | 0.403 | 19.62% |
| Riley Greene | 4 | 74.7 | 7.2 | 26.2 | 0.131 | 0.262 | 0.348 | 0.827 | 74 | 24 | 0.366 | 24.37% |
| Jackson Chourio | 6 | 73.1 | 7.4 | 24.5 | 0.138 | 0.275 | 0.327 | 0.791 | 79 | 21 | 0.351 | 24.41% |

Table 3: Season statistics of 6 batters that had Cluster 3 swings as their second most used swing type.

| Batter | Main Cluster | | | | | Cluster 3 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | wOBA | Whiff% | Blast Swing% | Barrel Swing% | Count | wOBA | Whiff% | Blast Swing% | Barrel Swing% | Count |
| Bryce Harper | 0.325 | 19.23% | 8.97% | 2.56% | 312 | 0.520 | 14.07% | 22.22% | 8.52% | 270 |
| Aaron Judge | 0.435 | 37.03% | 11.92% | 6.69% | 478 | 0.634 | 17.56% | 25.61% | 15.12% | 410 |
| Anthony Santander | 0.379 | 11.86% | 10.28% | 5.53% | 253 | 0.493 | 10.61% | 18.37% | 10.20% | 245 |
| Bobby Witt Jr. | 0.356 | 17.69% | 12.60% | 5.09% | 373 | 0.499 | 13.20% | 24.63% | 11.14% | 341 |
| Riley Greene | 0.341 | 17.04% | 9.00% | 2.89% | 311 | 0.457 | 18.78% | 20.00% | 10.20% | 245 |
| Jackson Chourio | 0.357 | 12.38% | 14.29% | 4.29% | 210 | 0.465 | 12.83% | 20.86% | 6.95% | 187 |

Table 4: Comparison of how the 6 batters performed in their main cluster (listed above) versus how they performed with Cluster 3 swings.

**Cluster 4:** Offering a similar power profile to Cluster 3 in rotational acceleration, its wOBA of .301 and Blast Swing% of 9.4%, demonstrated moderate effectiveness in generating runs. However, its slightly higher Whiff% (22.3%) and lower Squared-Up Swing% (22.9%) indicated that it lacked the versatility and consistency of Cluster 3 and Cluster 6.

**Cluster 5**: Despite a low Whiff% (16.56%) and moderate Squared Up Swing% (22.1%), these swings simply didn't produce enough bat speed to do any damage, as evidenced by its low wOBA (.250). Furthermore, Clusters 3 and 6 achieve higher wOBAs with lower Whiff%.

**Cluster 6**: With a low Whiff% (12.45%) and a solid Squared-Up% (31.82%), it ensured steady offensive output by maintaining contact and minimizing strikeouts. Cluster 6 excelled in high-pressure situations, particularly in two-strike counts and late innings, where reliable ball placement and consistent contact are paramount.

**Cluster 7**: Its low Whiff% (10.59%) and Squared-Up% (28.84%) highlighted its reliability in generating productive at-bats. While Cluster 7 didn't produce as much power as other clusters, its focus on contact made it effective for situational hitting, particularly when needing to advance runners.

## Key Findings

Our analysis highlights the situational dependence of optimal swing mechanics, revealing that no single swing type proves universally effective across all scenarios. Clusters 3 and 6 emerged as standout performers, offering complementary strengths that align with distinct game situations.

The high rotational acceleration and exceptional power metrics of Cluster 3 make it the most impactful swing type for producing game-altering hits. Its ability to generate high-quality contact while minimizing ground balls emphasizes its value in scenarios requiring offensive explosiveness. Moreover, its performance was particularly pronounced in close, late-game situations, where its efficiency in generating extra-base hits and driving in runs significantly shifted win probabilities. This swing type is best to apply in middle to late innings when teams seek to secure or expand their lead.

Cluster 6 provides a counterbalance with its focus on consistency and adaptability. By maintaining a low Whiff% and delivering a high Squared-Up%, it ensured steady performance in scenarios requiring reliability over explosiveness. Cluster 6 dominated in two-strike counts and high-leverage moments, offering a strategic tool for mitigating risks while advancing runners. This versatility underscores its importance in sustaining offensive momentum and capitalizing on situational opportunities.

We also found that moving runners over at the expense of an out can be beneficial, but only with no outs. Otherwise, the expected number of RBIs stays relatively similar even with the extra 90 feet. In Table 5, we can see that moving the runner over from second to third with no one out significantly increases the expected number of RBIs despite the extra out. However, when there's already 1 out, this tactic is less valuable. This is possibly due to the fact that sends from second become more aggressive with 2 outs.

Our study leveraged newly introduced Statcast metrics to provide a granular understanding of swing mechanics. By implementing k-means clustering to classify distinct swing types and running an XGBoost model to predict change in win probability, we deliver a robust methodological framework that enhances the validity and applicability of our findings.

These insights offer actionable strategies for coaches, scouts, and player development staff. By identifying each player's dominant swing cluster, teams can optimize lineups, implement strategic in-game adjustments, and design tailored training regimens that emphasize the strengths of individual players' swing types. The integration of advanced metrics and machine learning sets our study apart, offering a comprehensive toolkit for refining hitting strategies grounded in empirical data. Clusters 3 and 6 demonstrate the strategic depth of situational hitting, providing teams with the tools to maximize offensive output while minimizing risks.

| Cluster | Base Sit. | 0 Outs | 1 Out | 2 Outs |
|---|---|---|---|---|
| 1 | 3__ | 0.090 | 0.164 | 0.056 |
| 2 | 3__ | 0.305 | 0.229 | 0.134 |
| 3 | 3__ | 0.329 | 0.366 | 0.196 |
| 4 | 3__ | 0.377 | 0.269 | 0.132 |
| 5 | 3__ | 0.207 | 0.262 | 0.099 |
| 6 | 3__ | 0.390 | 0.374 | 0.154 |
| 7 | 3__ | 0.268 | 0.320 | 0.128 |
| 1 | _2_ | 0.023 | 0.040 | 0.055 |
| 2 | _2_ | 0.073 | 0.085 | 0.126 |
| 3 | _2_ | 0.193 | 0.180 | 0.195 |
| 4 | _2_ | 0.097 | 0.083 | 0.114 |
| 5 | _2_ | 0.050 | 0.079 | 0.108 |
| 6 | _2_ | 0.118 | 0.121 | 0.163 |
| 7 | _2_ | 0.069 | 0.099 | 0.100 |

Table 5: Displays the expected number of RBIs on each swing based on base situation, number of outs, and swing cluster. Here, we can see that moving the runner over to third may be valuable with no outs, but is not advisable if an out has already been made in the inning.

## Room for Improvement

While our analysis offers valuable insights towards the relationship between swing mechanics and situational performance, there are several opportunities for further refinement and exploration. First, something that our model doesn't account for is the momentum of the game. Earning just one run can help kickstart a big inning. Our analysis doesn't take this into account, but could be a useful consideration in future work.

Next, improving the clustering methodology by using dimensionality reduction techniques like t-SNE or PCA could uncover more nuanced swing patterns, while exploring hierarchical clustering might capture finer variations within swing types. Furthermore, integrating pitcher attributes (e.g., velocity, spin rate, pitch usage), defensive positioning, and baserunners' speed would provide a more comprehensive understanding of situational hitting and enhance our matrices by accounting for both offensive and defensive factors.

In addition, tracking how players' swings evolve over multiple seasons could reveal valuable understanding of their adaptability and performance. It would show how well players can adjust by switching between different swing styles to meet the demands of the game as they change. After the 2025 season, we could reevaluate our analysis and see if batters remain in their clusters or end up using a different swing type. We could also make a win percentage matrix with another season of data to discover a more meaningful pattern.

Lastly, expanding the predictive models to include psychological and environmental factors, such as confidence or game-time conditions, and exploring neural network architectures could further enhance the accuracy and depth of our analyses. By addressing these areas, future research can deepen our understanding of optimal swing strategies, providing even more robust tools for coaches, scouts, and player development staff to make data-driven decisions that maximize team success.

Works Cited


Agrawal S (2013). _Mini-Batch-K-Means_. <https://github.com/siddharth-agrawal/Mini-Batch-K-Means>.

Auguie B (2017). _gridExtra: Miscellaneous Functions for "Grid" Graphics_. R package version 2.3, <https://CRAN.R-project.org/package=gridExtra>.

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y, Yuan J (2024). _xgboost: Extreme Gradient Boosting_. R package version 1.7.7.1, <https://CRAN.R-project.org/package=xgboost>.

Dueck D (2009). _Affinity Propagation: Clustering Data by Passing Messages_. Ph.D. thesis, University of Toronto. <https://hdl.handle.net/1807/17755>.

Frey B, Dueck D (2007). “Clustering by Passing Messages Between Data Points.” _Science_, *315*, 972-976. doi:10.1126/science.1136800 <https://doi.org/10.1126/science.1136800>.

Frick H, Chow F, Kuhn M, Mahoney M, Silge J, Wickham H (2024). _rsample: General Resampling Infrastructure_. R package version 1.2.1, <https://CRAN.R-project.org/package=rsample>.

Kaplan J (2024). _fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables_. R package version 1.7.4, <https://CRAN.R-project.org/package=fastDummies>.

Kuhn M, Vaughan D, Hvitfeldt E (2024). _yardstick: Tidy Characterizations of Model Performance_. R package version 1.3.1, <https://CRAN.R-project.org/package=yardstick>.

Lambert, Jack. “Using MLB Bat Tracking Data to Better Understand Swings.” *Driveline Baseball*, 1 Aug. 2024, www.drivelinebaseball.com/2024/07/using-mlb-bat-tracking-data-to-better-understand-swings/.

Mouselimis L (2024). _ClusterR: Gaussian Mixture Models, K-Means, Mini-Batch-Kmeans, K-Medoids and Affinity Propagation Clustering_. R package version 1.3.3, <https://CRAN.R-project.org/package=ClusterR>.

Petti B, Gilani S (2024). _baseballr: Acquiring and Analyzing Baseball Data_. R package version 1.6.0, <https://CRAN.R-project.org/package=baseballr>.

Pham D, Dimov S, Nguyen C (2004). “Selection of K in K-means clustering.” _Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science_. doi:10.1243/095440605X8298 <https://doi.org/10.1243/095440605X8298>,

Sanderson C, Curtin R (2016). “Armadillo: a template-based C++ library for linear algebra.” _Journal of Open Source Software_, *1*, 26. doi:10.21105/joss.00026 <https://doi.org/10.21105/joss.00026>.

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Richardson N, Cook I, Crane N, Dunnington D, François R, Keane J, Moldovan-Grünfeld D, Ooms J, Wujciak-Jens J, Apache Arrow (2024). _arrow: Integration to 'Apache Arrow'_. R package version 16.1.0, <https://CRAN.R-project.org/package=arrow>.

Sanderson C, Curtin R (2018). "A User-Friendly Hybrid Sparse Matrix Class in C++." _Lecture Notes in Computer Science (LNCS)_, *10931*, 422-430.doi:10.1007/978-3-319-96418-8_50 <https://doi.org/10.1007/978-3-319-96418-8_50>.

Sculley D (2010). "Web-scale k-means clustering." _ACM Digital Library_. doi:10.1145/1772690.1772862 <https://doi.org/10.1145/1772690.1772862>.

Struyf A, Hubert M, Rousseeuw P (1997). "Clustering in an Object-Oriented Environment." _Journal of Statistical Software_. doi:10.18637/jss.v001.i04 <https://doi.org/10.18637/jss.v001.i04>.

Tango, Tom. "Statcast Lab: Collisions and the Perfect Swing." *Tangotiger Blog*, 22 June 2022, tangotiger.com/index.php/site/article/statcast-lab-collisions-and-the-perfect-swing.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." _Journal of Open Source Software_, *4*(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686>.

Wilke C (2024). _ggridges: Ridgeline Plots in 'ggplot2'_. R package version 0.5.6, <https://CRAN.R-project.org/package=ggridges>.

Zeileis A, Fisher JC, Hornik K, Ihaka R, McWhite CD, Murrell P, Stauffer R, Wilke CO (2020). "colorspace: A Toolbox for Manipulating and Assessing Colors and Palettes." _Journal of Statistical Software_, *96*(1), 1-49. doi:10.18637/jss.v096.i01 <https://doi.org/10.18637/jss.v096.i01>